National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Architecture for Distributed Earth Science Data Analysis

COVERAGE Architecture | CEOS 2018 SIT Technical Workshop

Thomas Huang

Technical Group Supervisor, Computer Science for Data-Intensive Applications

Instrument Software and Science Data Systems

thomas.huang@jpl.caltech.edu

Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109-8099, U.S.A.

Data Centers

- **Mainly focus on archives and distributions**
- **With additional services**
  - Better searches – faceted, spatial, keyword, ranking, etc.
  - Data subsetting – home grown, OPeNDAP, Webification, etc.
  - Visualization – visual discovery, PO.DAAC's SOTO, NASA Worldview, etc.
- **Limitations**
  - Little to no interoperability between tools and services: metadata standard, keyword, spatial coverage (0-360 or -180..180), temporal representation, etc.
  - Making sure the most relevant measurements return first
  - Visualization is nice, but it doesn't provide enough information about the event/phenomenon captured in the image.
  - With large amount of observational data, data centers need to do more than just storing bits
    - "Is the red blob in the middle of Pacific normal this time of the year?"
    - "Any relevant news and publications relate to what I am looking at?"
    - "What other measurements, phenomena, news, publications relate to the period and location I am looking at?"
    - "I can see the observation from satellite, are there any relevant in situ data I can look at?"

Background

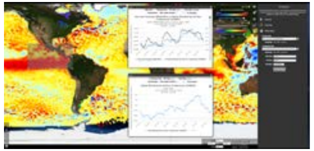
- **With large amount of observational and modeling data, finding and downloading is becoming inefficient**
- **Reality with large amount of observational and modeling data**
  - Downloading to local machine is becoming inefficient
  - Search has gotten a lot faster. Too many matches.
  - Finding the relevant measurement has becoming a very time consuming process ”*Which SST dataset I should use?*”
  - Analyze decades of regional measurement is labor-intensive and costly
- **Increasing “big data” era is driving needs to**
  - Scale computational and data infrastructures
  - Support new methods for deriving scientific inferences
  - Shift towards integrated data analytics
  - Apply computational and data science across the lifecycle
- **Scalable Data Management**
  - Capture well-architected and curated data repositories based on well-defined data/information architectures
  - Architecting automated pipelines for data capture
- **Scalable Data Analytics**
  - Access and integration of highly distributed, heterogeneous data
  - Novel statistical approaches for data integration and fusion
  - Computation applied at the data sources
  - Algorithms for identifying and extracting interesting features and patterns

National Aeronautics and Space Administration

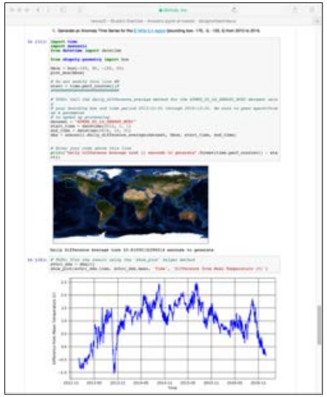
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Enabling Next Generation of Ocean Science Tools and Services

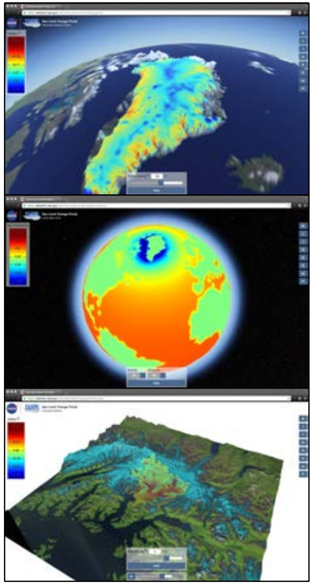
NASA Sea Level Change Portal



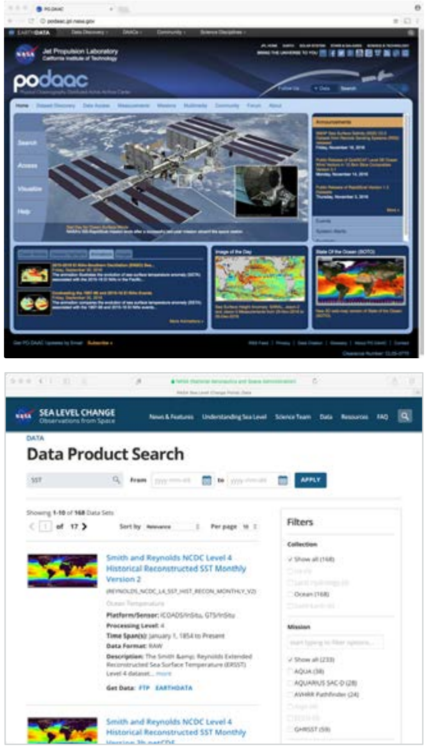
Jupyter Notebook - Interactive Workbench



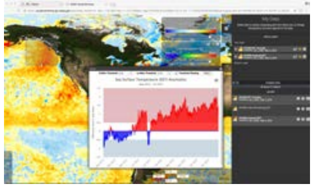
Model Simulations



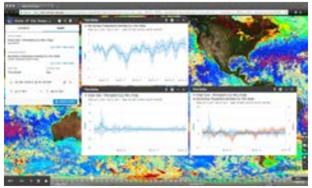
Integrated Search and Discovery



Oceanographic Anomaly Detection

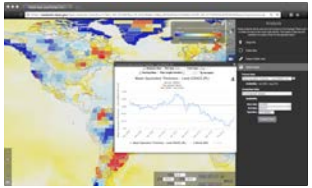


PO.DAAC State Of The Ocean

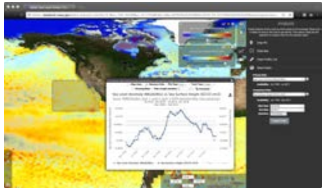


Mobile Analysis

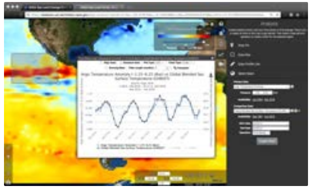
Hydrological Basin Analysis



In Situ Data Analysis



Model - Observation Comparison



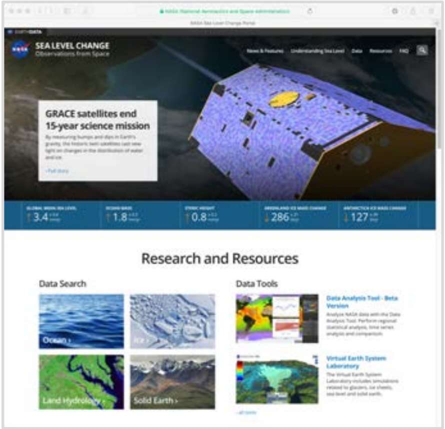
Reference Architecture – NASA’s Sea Level Change Portal

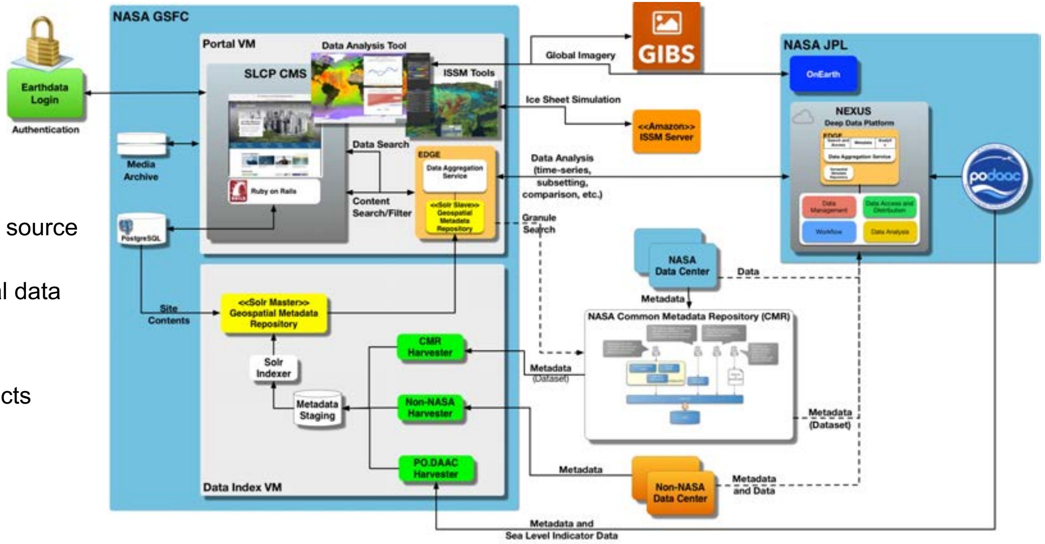
Goal for the NASA Sea Level Change Team

- Determine how much will sea level rise by [2100]?
- What are the key sensitivities?
- Where are the key uncertainties? Observables? Model Improvements

Goals for the NASA Sea Level Change Portal

- Provide scientists and the general public with a “one-stop” source for current sea level change information and data
- Provide interactive tools for analyzing and viewing regional data
- Provide virtual dashboard for sea level indicators
- Provide latest news, quarterly report, and publications
- Provide ongoing updates through a suite of editorial products



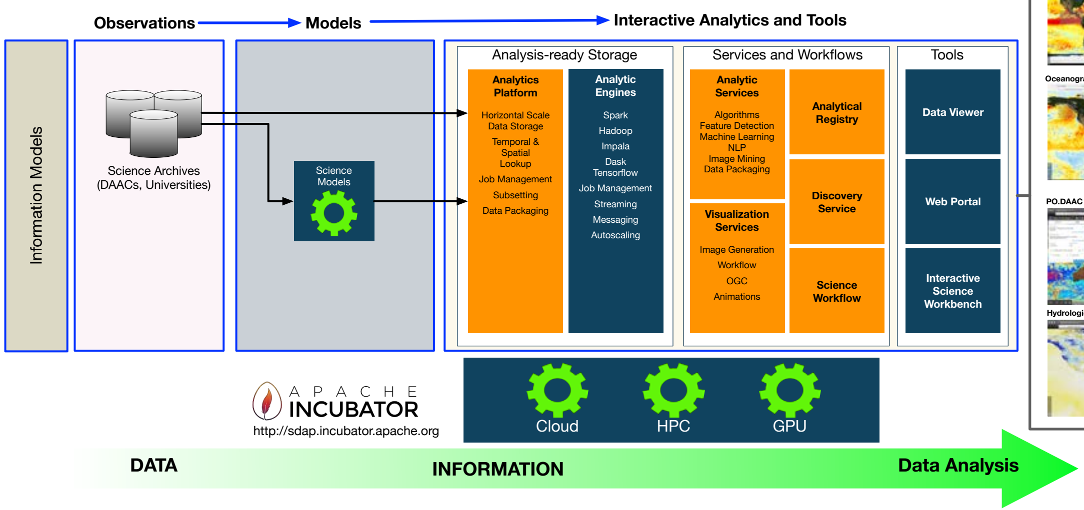


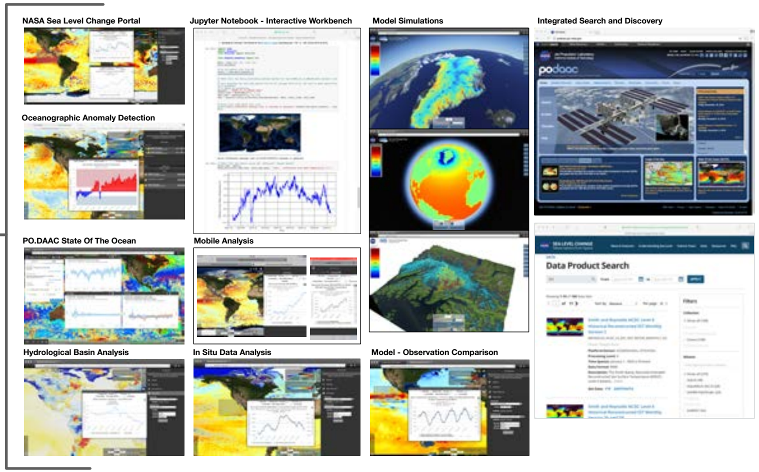
NASA’s Sea Level Change Portal System Architecture

https://sealevel.nasa.gov

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Integrated Ocean Science Data Analytics Platform



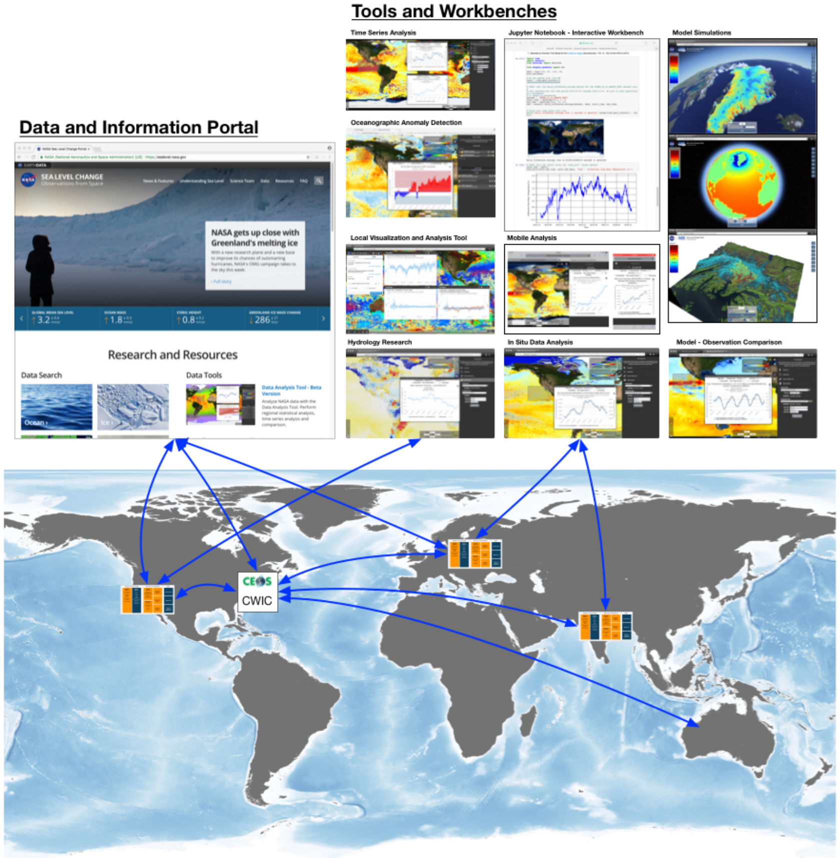


Systematic analysis of massive data

- **Integrated Ocean Science Data Analytics Platform**: an environment for conducting a Ocean Science investigation
  - Enables the confluence of resources for that investigation
  - Tailored to the individual study area (physical ocean, sea level, etc.)
- Harmonizes data, tools and computational resources to permit the ocean research community to focus on the investigation
- Scale computational and data infrastructures
- Shift towards integrated data analytics
- Algorithms for identifying and extracting interesting features and patterns

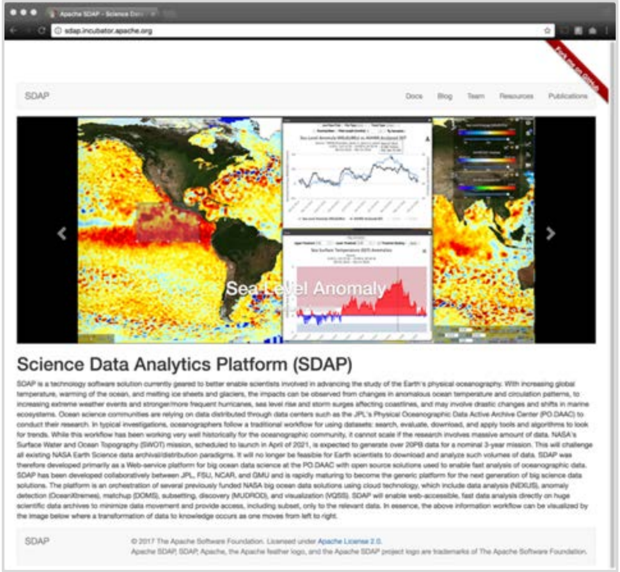
Architecture for Distributed Data System and Analysis

- A common service registry, perhaps CWIC?
- Putting value-added services next to the data to eliminate unnecessary data movement
- Avoid data replication
- Public accessible RESTful analytic APIs where computation is next to the data
- Analytic engine infused and managed by the data centers perhaps on the Cloud
- Researchers can perform multi-variable analysis using any web-enabled devices without having to download files
- Reduce unnecessary data movement and egress charges
- An architecture to enable next generation of scientific applications



Open Source

- Technology sharing through Free and Open Source Software (FOSS)
- Further technology evolution that is restricted by projects / missions
- **Science Data Analytic Platform (SDAP)**, the implementation of **OceanWorks**, in **Apache Incubator**
    - Cloud platform
    - Analyzing satellite and model data
    - In situ data analysis and colocation with satellite measurements
    - Fast data subsetting
    - Mining of user interactions and data to enable discovery and recommendations
    - Streamline deployment through container technology

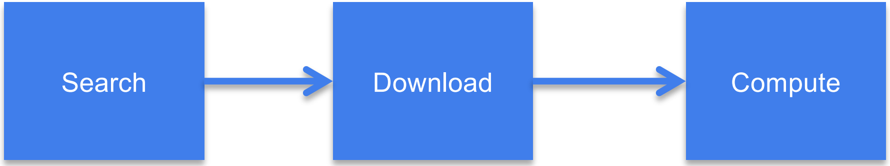


http://sdap.incubator.apache.org

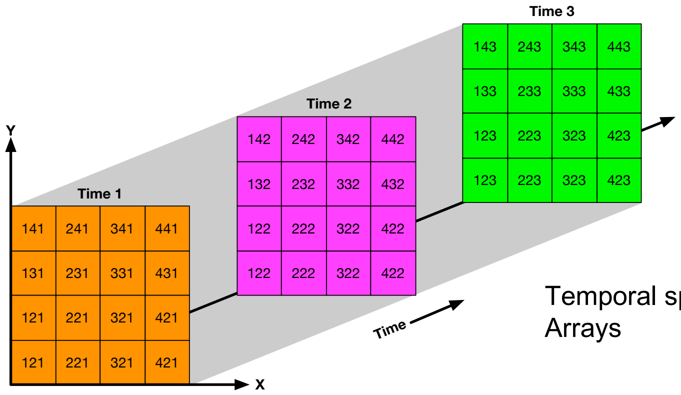
APACHE INCUBATOR

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Traditional Method for Analyze Satellite Measurements



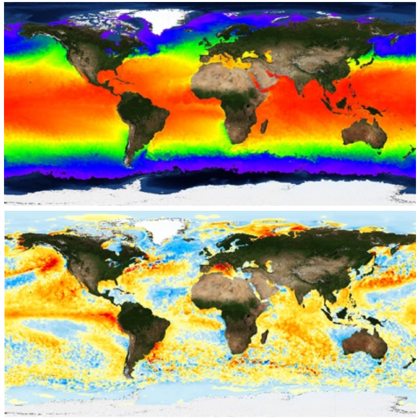
- Depending on the data volume (size and number of files)
- It could take many hours of download – (e.g. 10yr of observational data could yield thousands of files)
- It could take many hours of computation
- It requires expensive local computing resource (CPU + RAM + Storage)
- After result is produced, purge downloaded files

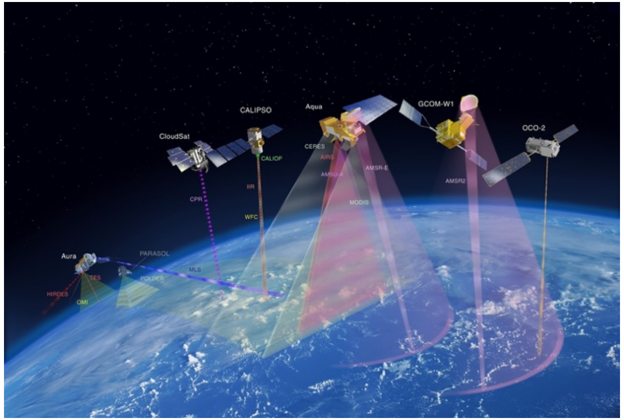


Temporal spatial Arrays

Observation

- Traditional methods for data analysis (time-series, distribution, climatology generation) can't scale to handle large volume, high-resolution data. They perform poorly
- Performance suffers when involve large files and/or large collection of files
- A high-performance data analysis solution must be free from file I/O bottleneck



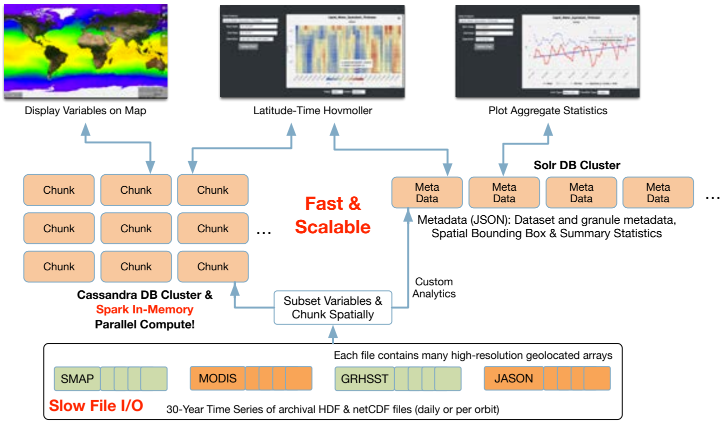


National Aeronautics and Space Administration

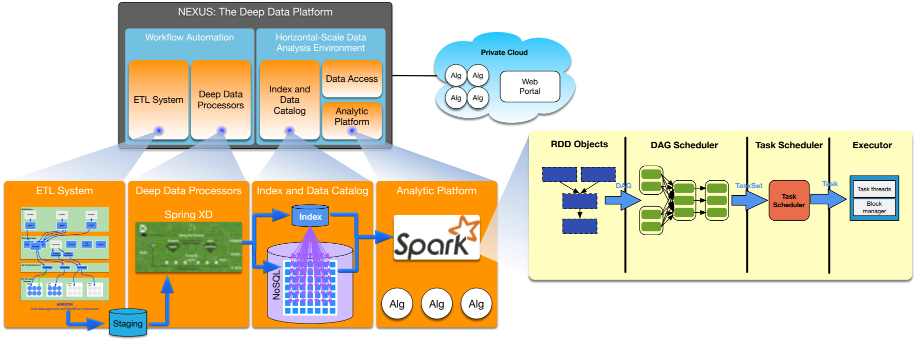
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

NEXUS: Scalable Data Analytic Solution

- NEXUS is a data-intensive analysis solution using a new approach for handling science data to enable large-scale data analysis
- Streaming architecture for horizontal scale data ingestion
- Scales horizontally to handle massive amount of data in parallel
- Provides high-performance geospatial and indexed search solution
- Provides tiled data storage architecture to eliminate file I/O overhead
- A growing collection of science analysis webservices using Apache Spark: parallel compute, in-memory map-reduce framework
- Pre-Chunk and Summarize Key Variables
    - Easy statistics instantly (milliseconds)
    - Harder statistics on-demand using Spark (in seconds)
    - Visualize original data (layers) on a map quickly (Cassandra store)
- **Algorithms** – Time Series | Latitude/Time Hovmöller| Longitude/Time Hovmöller| Latitude/Longitude Time Average | Area Averaged Time Series | Time Averaged Map | Climatological Map | Correlation Map | Daily Difference Average



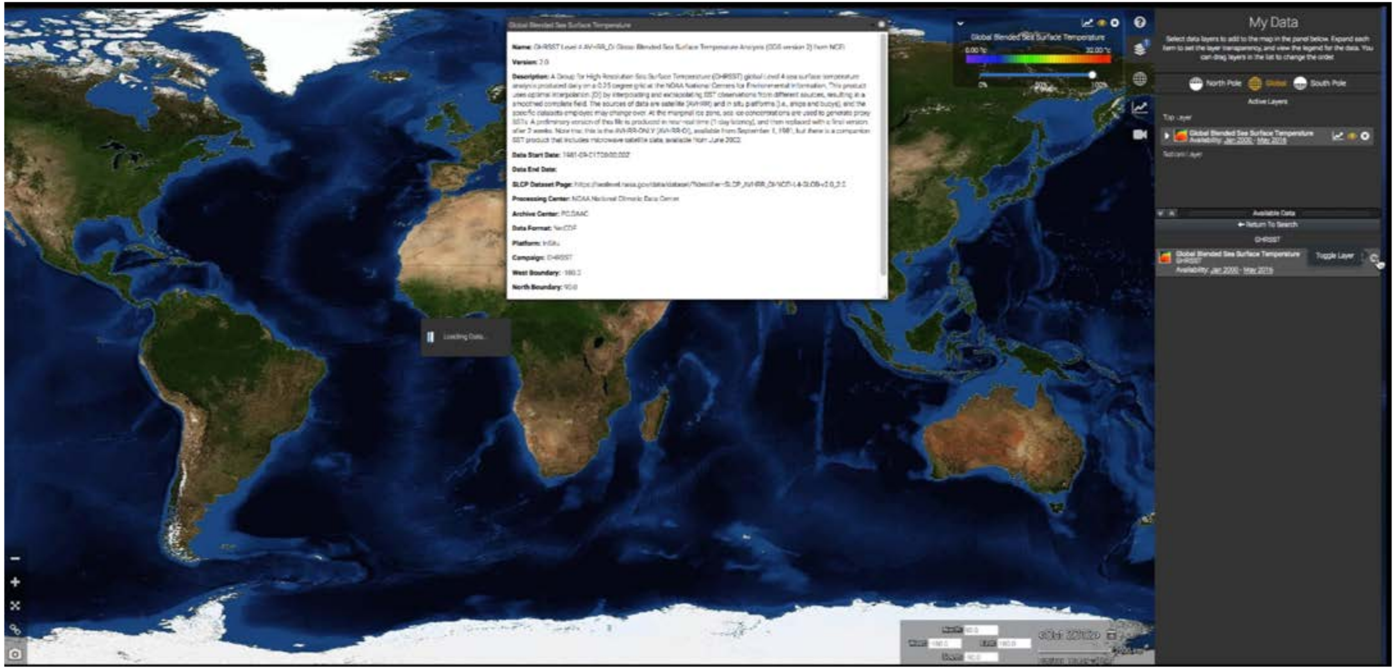
Two-Database Architecture



National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Analyze Sea Level On-The-Fly *https://sealevel.nasa.gov*



Sea Level Change - Data Analysis Tool

Visualizations | Hydrological Basins | Time Series | Deseason | Data Comparison | Scatter Plot | Latitude/Time Hovmöller | Etc.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

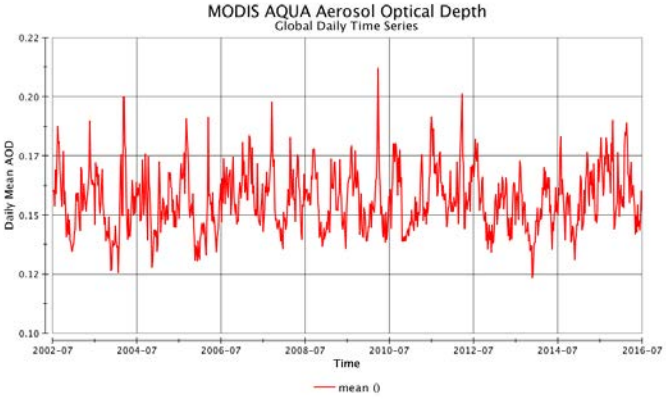
NEXUS Performance: GIOVANNI(v4) vs. Custom Spark vs. AWS EMR

Dataset: MODIS AQUA Daily
Name: Aerosol Optical Depth 550 nm (Dark Target) (MYD08_D3v6)
File Count: 5106
Volume: 2.6GB
Time Coverage: July 4, 2002 – July 3, 2016

Giovanni: A web-based application for visualize, analyze, and access vast amounts of Earth science remote sensing data without having to download the data.

- Represents current state of data analysis technology, by processing one file at a time
- Backed by the popular NCO library. Highly optimized C/C++ library

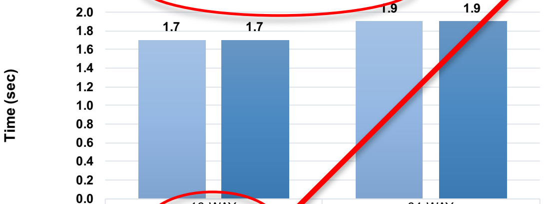
AWS EMR: Amazon’s provisioned MapReduce cluster



GIOVANNI: 20 min
NEXUS: 1.7 sec

Area Averaged Time Series on AWS - Boulder
July 4, 2002 - July 3, 2016
NEXUS Performance

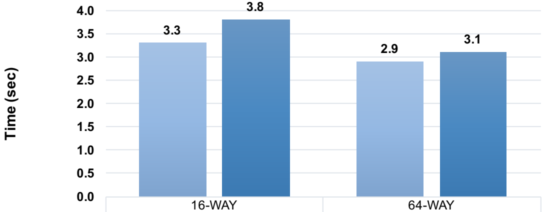
Custom Spark vs. AWS EMR
Ref. Speed - Giovanni: 1140.22 sec



	16-WAY	64-WAY
Custom Spark	1.7	1.9
AWS EMR	1.7	1.9

Area Averaged Time Series on AWS - Colorado
July 4, 2002 - July 3, 2016
NEXUS Performance

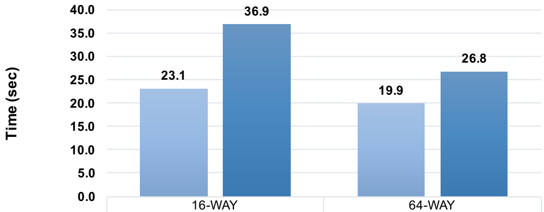
Custom Spark vs. AWS EMR
Ref. Speed - Giovanni: 1150.6 sec



	16-WAY	64-WAY
Custom Spark	3.3	2.9
AWS EMR	3.8	3.1

Area Averaged Time Series on AWS - Global
July 4, 2002 - July 3, 2016
NEXUS Performance

Custom Spark vs. AWS EMR
Ref. Speed - Giovanni: 1366.84 sec



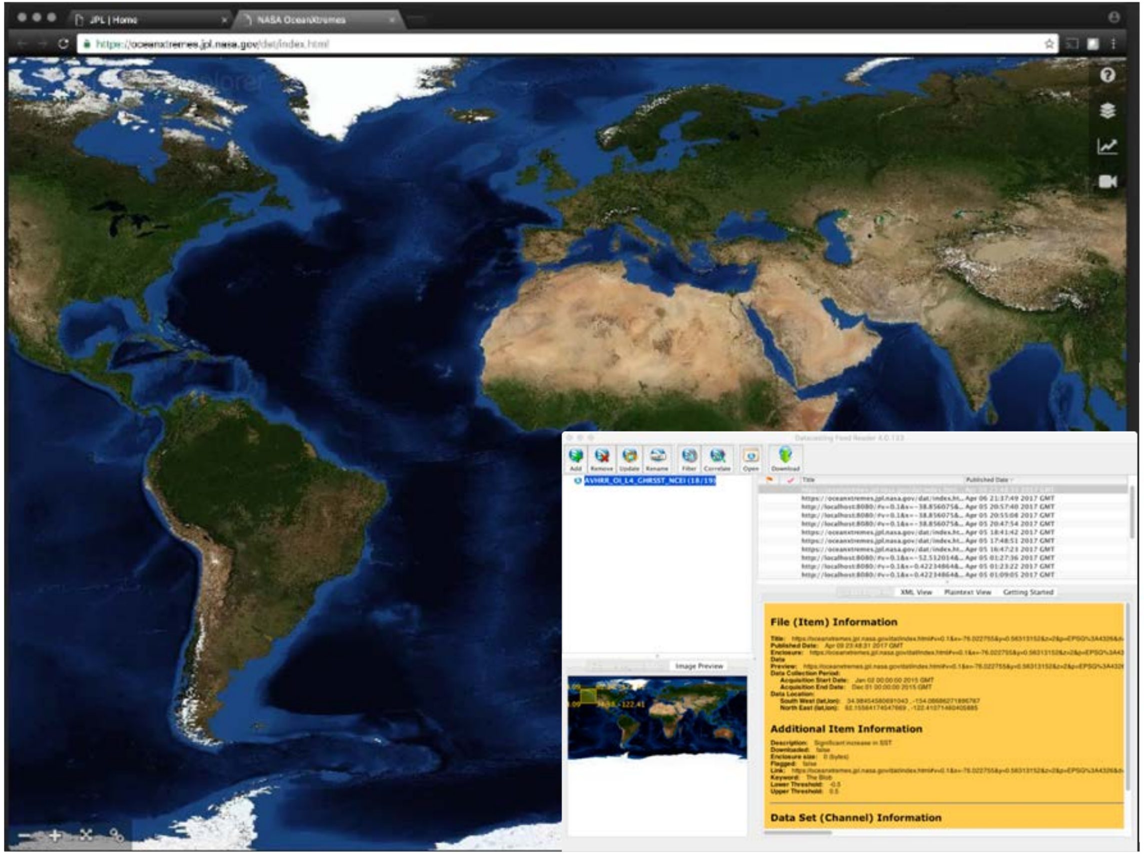
	16-WAY	64-WAY
Custom Spark	23.1	19.9
AWS EMR	36.9	26.8

Algorithm execution time. Excludes Giovanni’s data scrubbing processing time

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Analyze Ocean Anomaly – “The Blob”



- **Visualize** parameter
- **Compute** daily differences against climatology
- **Analyze** time series area averaged differences
- **Replay** the anomaly and visualize with other measurements
- **Document** the anomaly
- **Publish** the anomaly

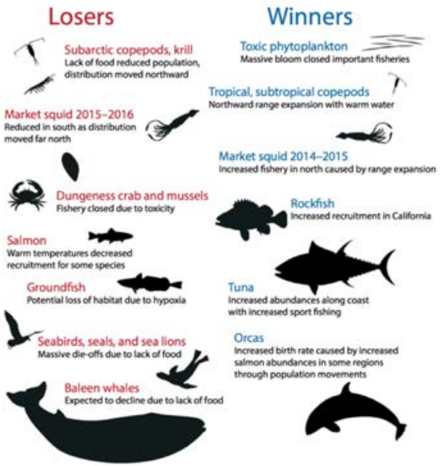
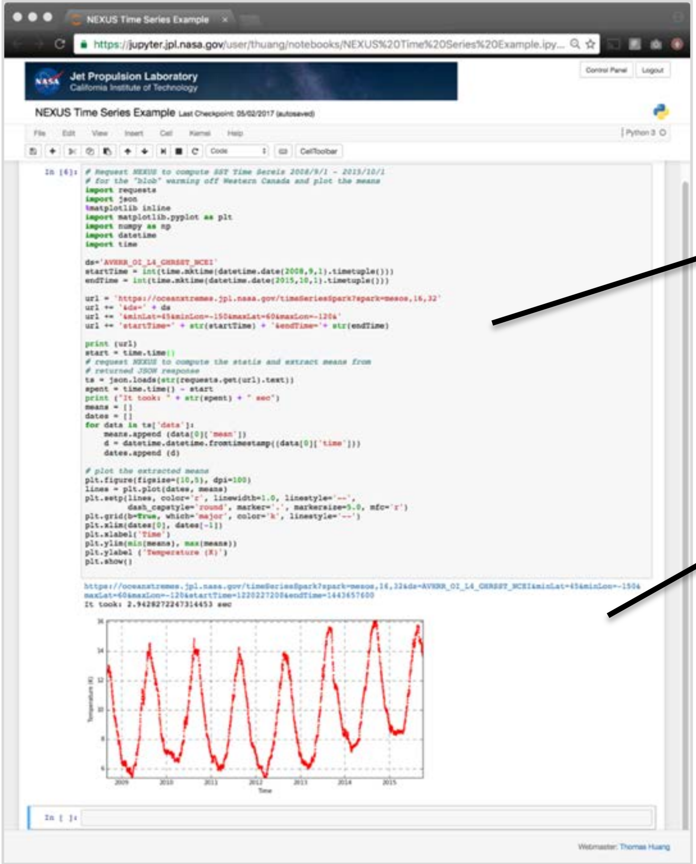


Figure from Cavole, L. M., et al. (2016). "Biological Impacts of the 2013–2015 Warm-Water Anomaly in the Northeast Pacific: Winners, Losers, and the Future." Oceanography 29.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Enable Science without File Download




```
# Request NEXUS to compute SST Time Series 2008/9/1 - 2015/10/1
# for the "blob" warming off Western Canada and plot the means
…
ds='AVHRR_OI_L4_GHRSST_NCEI'

url = … # construct the webservice URL request


# make request to NEXUS using URL request
# save JSON response in local variable
ts = json.loads(str(requests.get(url).text))

# extract dates and means from the response
means = []
dates = []
for data in ts['data']:
    means.append (data[0]['mean'])
    d = datetime.datetime.fromtimestamp((data[0]['time']))
    dates.append (d)

# plot the result
…
```

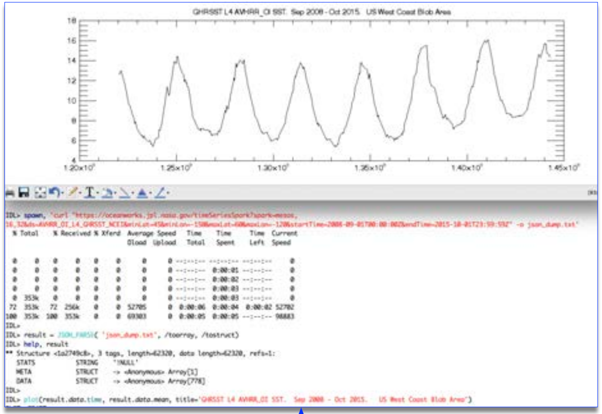
```
https://oceanxtremes.jpl.nasa.gov/timeSeriesSpark?spark=mesos,16,32&ds=AVHRR_OI_L4_GHRSST_NCEI&minLat=45&minLon=-150&maxLat=60&maxLon=-120&startTime=1220227200&endTime=1443657600

It took: 2.9428272247314453 sec
```

Using IDL with NEXUS

```
IDL> spawn, 'curl
"https://oceanworks.jpl.nasa.gov/timeSeriesSpark?spark=mesos,16,32&ds=AVHRR_OI_L4_GHRSST_NCEI&minLat=45&minLon=-150&maxLat=60&maxLon=-120&startTime=2008-09-01T00:00:00Z&endTime=2015-10-01T23:59:59Z" -o json_dump.txt'
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed

  0      0    0      0    0     0      0      0 --:--:-- --:--:-- --:--:--     0
  0      0    0      0    0     0      0      0 --:--:--  0:00:01 --:--:--     0
  0      0    0      0    0     0      0      0 --:--:--  0:00:02 --:--:--     0
  0      0    0      0    0     0      0      0 --:--:--  0:00:03 --:--:--     0
  0   353k    0      0    0     0      0      0 --:--:--  0:00:03 --:--:--     0
 72   353k   72   256k    0     0  52705      0  0:00:06  0:00:04  0:00:02 52702
100   353k  100   353k    0     0  69303      0  0:00:05  0:00:05 --:--:-- 98883
IDL>
IDL> result = JSON_PARSE( 'json_dump.txt', /toarray, /tostruct)
IDL> help, result
** Structure <1a2749c8>, 3 tags, length=62320, data length=62320, refs=1:
   STATS           STRING    '!NULL'
   META            STRUCT    -> <Anonymous> Array[1]
   DATA            STRUCT    -> <Anonymous> Array[778]
IDL>
IDL> plot(result.data.time, result.data.mean, title='GHRSST L4 AVHRR_OI SST.  Sep 2008
- Oct 2015.   US West Coast Blob Area')
PLOT <29457>
```



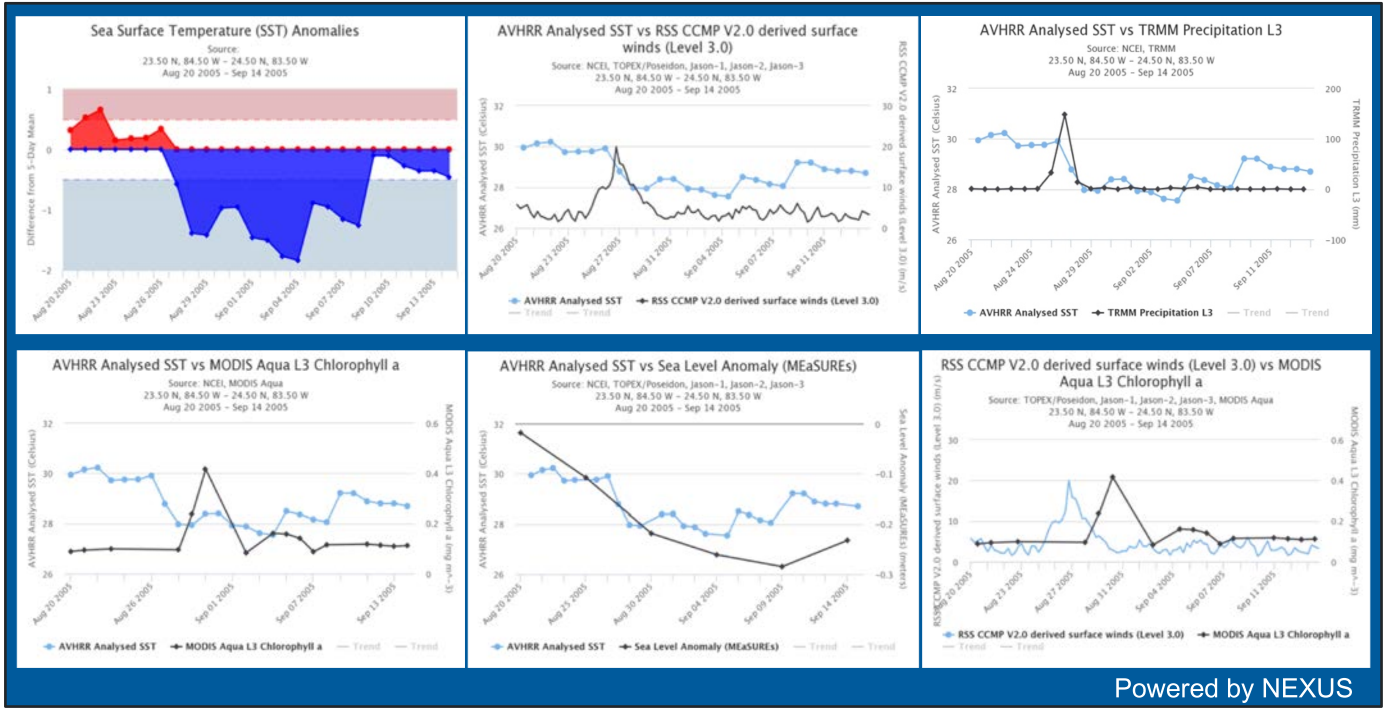
Credit: Ed Armstrong

Jun. 05, 2018

National Aeronautics and Space Administration

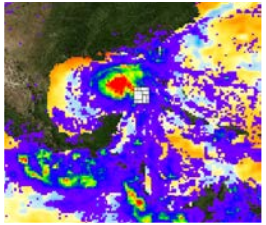
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Hurricane Katrina Study



Hurricane Katrina passed to the southwest of Florida on Aug 27, 2005. The ocean response in a 1 x 1 deg region is captured by a number of satellites. The initial ocean response was an immediate cooling of the surface waters by 2 ℃ that lingers for several days. Following this was a short intense ocean chlorophyll bloom a few days later. The ocean may have been "preconditioned' by a cool core eddy and low sea surface height.

The SST drop is correlated to both wind and precipitation data. The Chl-A data is lagged by about 3 days to the other observations like SST, wind and precipitation.

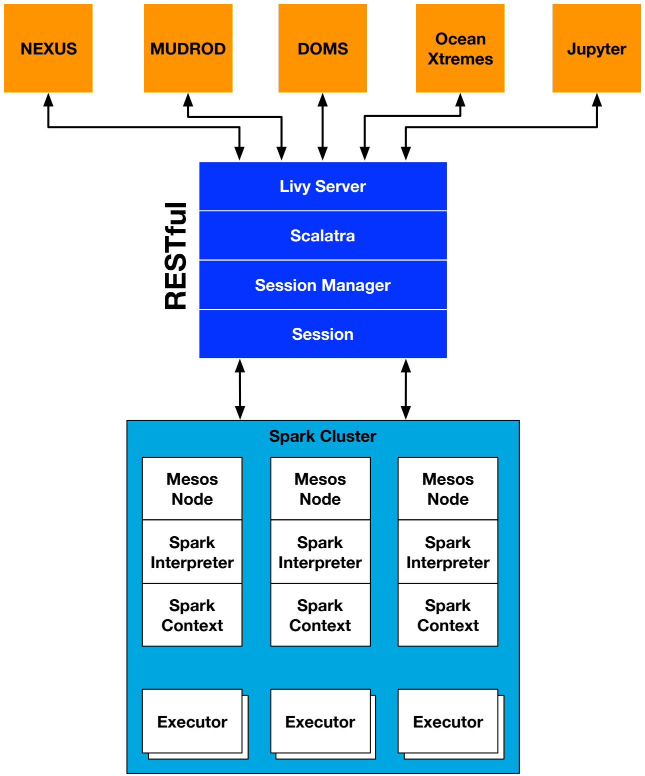


Hurricane Katrina TRMM overlay SST Anomaly

A study of a Hurricane Katrina–induced phytoplankton bloom using satellite observations and model simulations
Xiaoming Liu, Menghua Wang, and Wei Shi
JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 114, C03023, doi:10.1029/2008JC004934, 2009

Architecture for Apache Spark Integration and Sharing

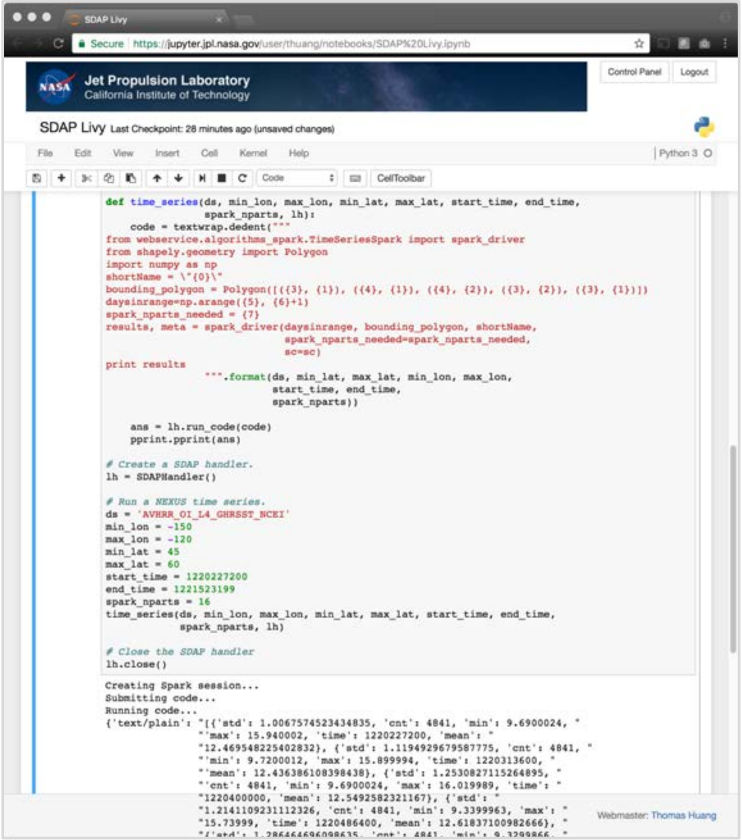
- Apache Spark has become the de facto framework for many data analytics problems. Spinning new Spark cluster for each service is undesirable
- Too many cluster and very costly, since Apache Spark recommends high memory machine instances
- Looking at the Amazon's EMR model. It is designed to be a job execution solution, and the jobs could from different applications
- Apache Livy provides a RESTful interface to Apache Spark cluster. It is a drop-in service to enable applications to interact with Spark cluster using RESTful API.
- Spark-enabled applications can use Livy to interface with a shared Spark cluster



Architecture for Sharing Spark Cluster

Push Python Code to Our Spark Cluster

- Using RESTful API, researchers can now push ad hoc map-reduce algorithm in Python to execute by our Spark cluster
- It provides a flexible environment for researchers to experiment with their algorithms and our data, without having to deal with the complexity of Cloud and job management



Analytic Storage

Blocked
Storage
(EBS)

v.s.

Object
Storage
(S3)

$0.045/GB-month
$47,186/PB-month

$0.021/GB-month
$22,020/PB-month

NEXUS supports 2 types of cloud storage

- <u>Blocked storage</u> (e.g. Amazon EBS), Attach to computing node. Generally faster
  - Cassandra and ScyllaDB
- <u>Object storage</u> (e.g. Amazon S3), Independent storage service. Highly scalable

For a data center, not all data need to be served on fast storage, Object storage provides a better, scalable alternative

Stream-based Ingestion Workflow Architecture

- Data streaming architecture
  - Applications are connected to form ingestion streams
  - Configurable to handle different datasets
    - Multiple tiling algorithms
    - Support L2 swath data
    - Support gridded data
  - Scalable across compute resources
  - Resilient to failure
  - ESDS-RFC-028v1.1

```
def filter_empty_tiles(self, tile):
  # Only supply data if there is actual values in the tile
  if tile.data.size - numpy.count_nonzero(numpy.isnan(tile.data)) > 0:
    yield tile.data
  else:
    print "Discarding data %s from %s because it is empty" % (tile.section_spec, tile.granule)
```

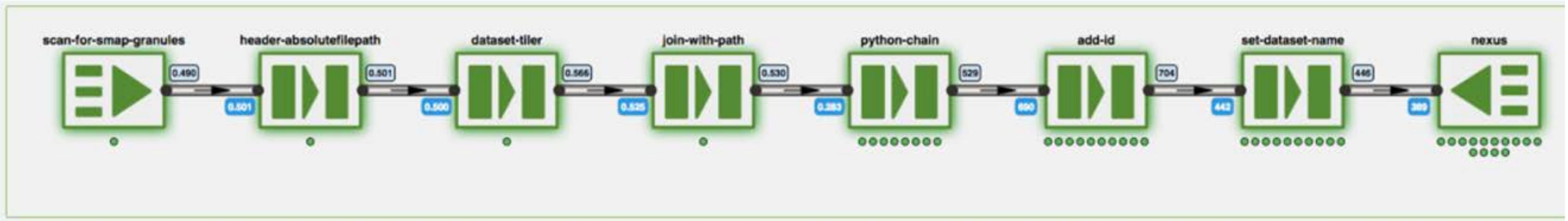
Pluggable validation checkers

```
def transform(self, tile):

  # Subtract 360 if the longitude is greater than 180
  tile.data.longitudes[longitudes > 180] -= 360

  yield tile.data
```

Data translation



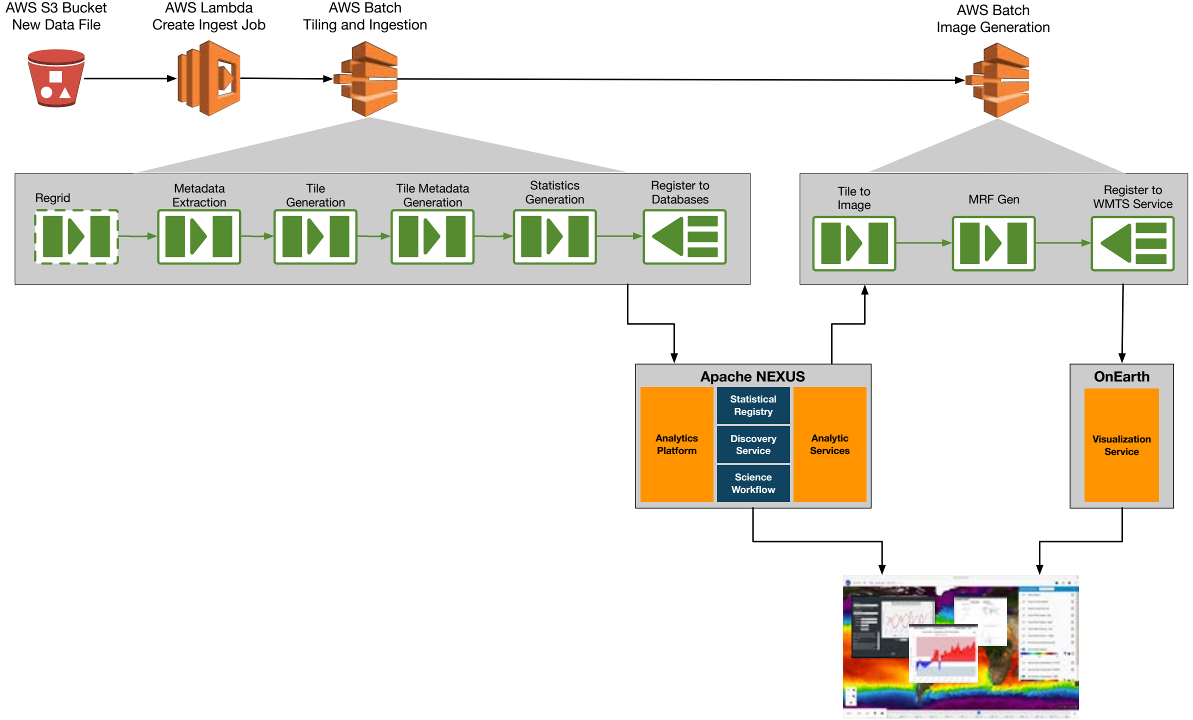
Stream for AVHRR_OI-NCEI-L4-GLOB-v2.0 Sea Surface Temperature

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

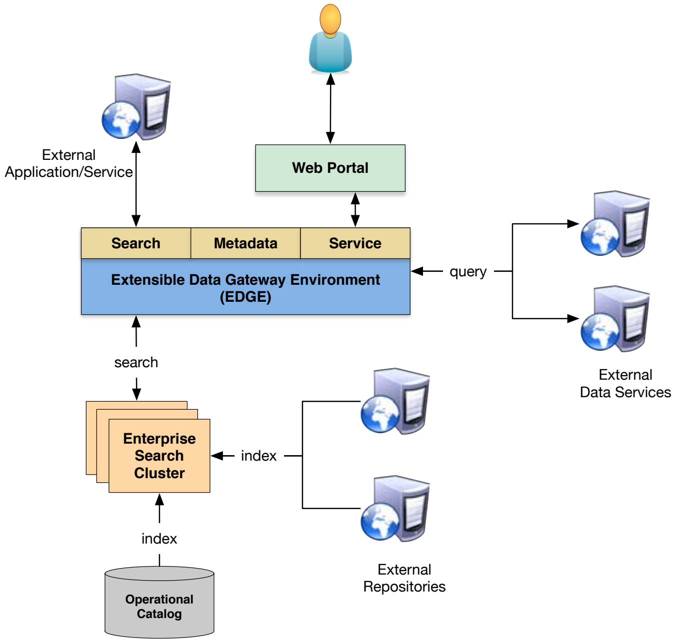
Severless Ingestion Workflow Architecture

- The cluster ingestion workflow requires several EC2 instances and attached EBS storage. Amazon bills active EC2 and storage even there is no data to be processed
- Cost-saving serverless architecture using AWS S3 Bucket trigging AWS Lambda job to create AWS Batch jobs



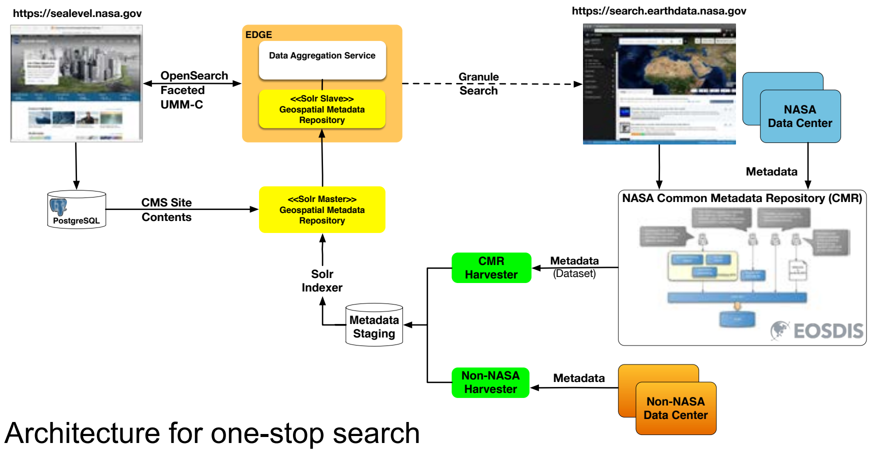
Extensible Data Gateway Environment (EDGE)

- Open Source high-performance geospatial data search and access
- Delivers sub-second search solution
- Implements the ESIP Federation's Discovery Specification (http://wiki.esipfed.org/index.php/Discovery_Cluster), which is a specialization of the OpenSearch (http://www.opensearch.org) standard (both XML and JSON)
- Platform to support multi-metadata standard specifications including ISO-19115, NASA UMM-C, NASA ECHO-10, NASA Global Change Master Directory (GCMD), Federation Geographic Data Committee (FGDC), and various domain-specific metadata standards
- Two main building blocks: data aggregation service and enterprise geospatial indexed search cluster
- Aggregation – provides a plugin approach to integrate with other external data repositories by proxying to other local/remote data services to reduce the number of interfaces a requestor has to access
- Enterprise geospatial indexed search cluster for fast lookup. Supports Apache Solr (and SolrCloud) and ElasticSearch
- Various production deployments including NASA Sea Level Change Portal, GRACE Web Portal, PO.DAAC, NASA ACCESS and AIST projects, and various Navel Research projects

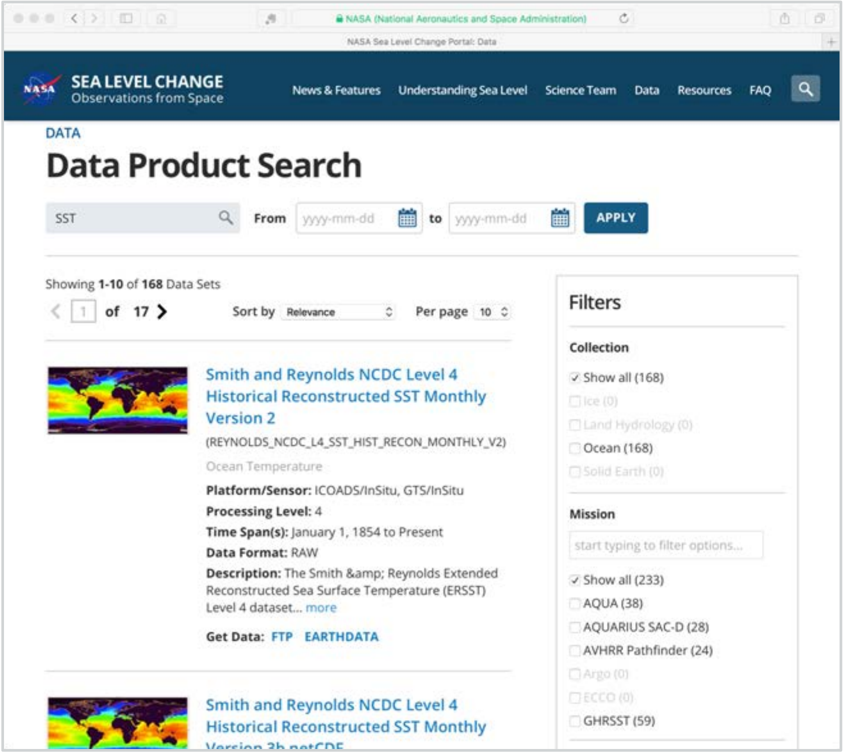


NASA Sea Level Change Portal's One-Stop Search

- Homogenize metadata acquired from different providers
- On-the-fly translation metadata and search results according to the NASA ECHO-10 and UMM-C specification
- Simplify web portal integration by providing one-stop search solution for all Sea Level artifacts – data, news, publications, and multi-media resources



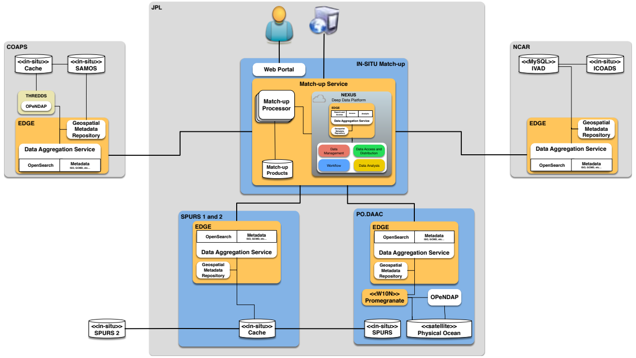
Architecture for one-stop search

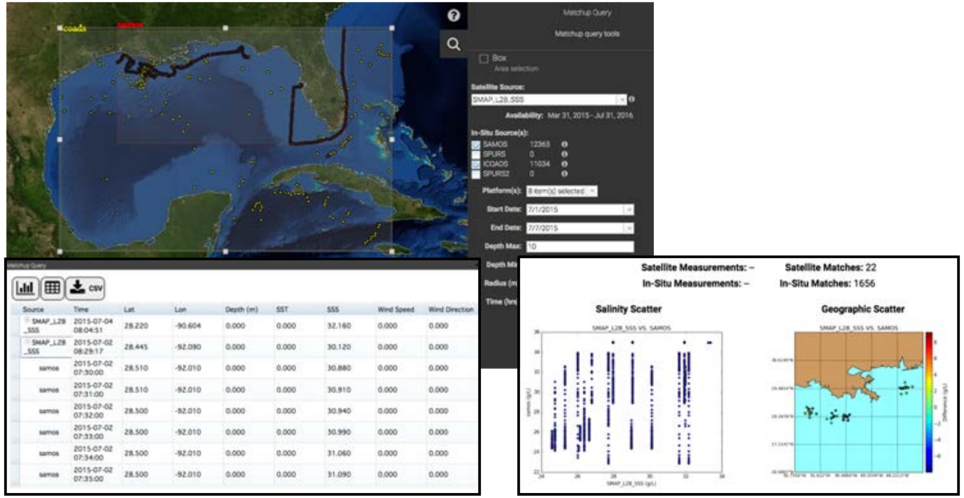


NASA Sea Level Change Portal's One-Stop Search

In Situ to Satellite Matchup

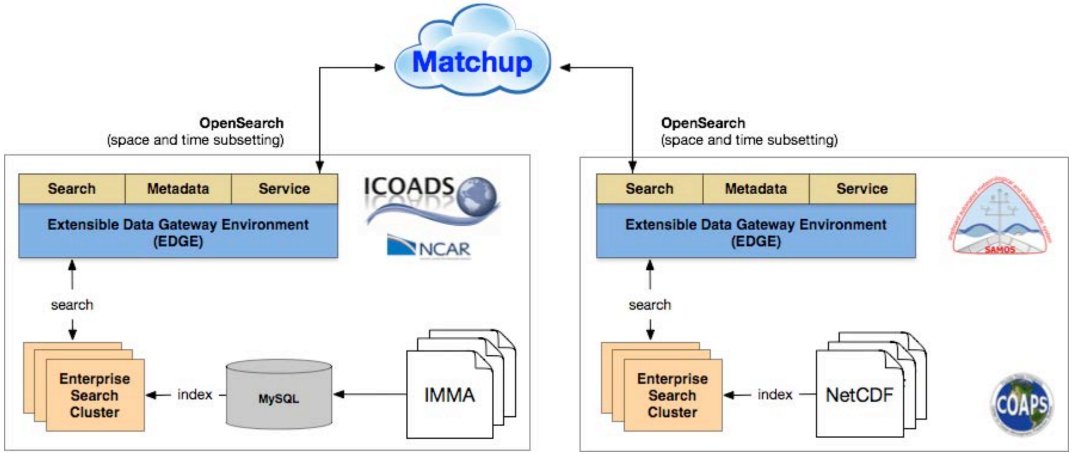
- Distributed Oceanographic Matchup Service (DOMS)
- Typically data matching is done using one-off programs developed at multiple institutions
- A primary advantage of DOMS is the reduction in duplicate development and man hours required to match satellite/in situ data
    - Removes the need for satellite and in situ data to be collocated on a single server
    - Systematically recreate matchups if either in situ or satellite products are re-processed (new versions), i.e., matchup archives are always up-to-date.
- In situ data nodes at JPL, NCAR, and FSU operational.
- Provides data querying, subset creation, match-up services, and file delivery operational.
- Plugin architecture for in situ data source using EDGE, a open source implementation of Open Search





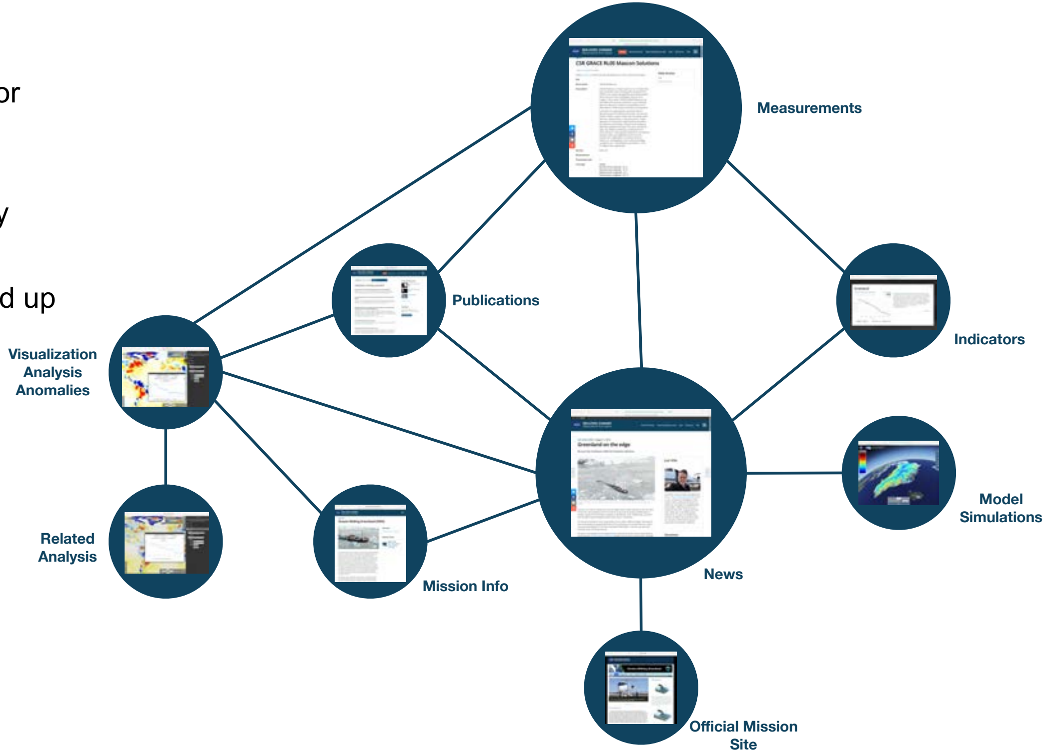
On-the-Fly Subsetting of In-Situ Measurements using OpenSearch

- Using OpenSearch as the standard interface to in-situ data repositories
- Enable distributed, federated search and data subsetting
- Subset in-situ data by time and space using OpenSearch
  - **ICOADS:** `'http://rda-data.ucar.edu:8890/ws/search/icoads?startTime=2012-08-01T00:00:00Z&endTime=2013-10-31T23:59:59Z&bbox=-45,15,-30,30'`
  - **SAMOS:** `'http://doms.coaps.fsu.edu/edge/samos?startTime=2012-08-01T00:00:00Z&endTime=2013-10-31T23:59:59Z&bbox=-45,15,-30,30'`
  - **SPURS-1:** `'https://doms.jpl.nasa.gov/spurs?startTime=201208-01T00:00:00Z&endTime=2013-10-31T23:59:59Z&bbox=-45,15,-30,30'`
  - **SPURS-2:** `'https://doms.jpl.nasa.gov/spurs2?startTime=2016-07-01T00:00:00Z&endTime=2016-07-31T23:59:59Z&bbox=-142,4,-112,24'`



Tackling Information Discovery

- **Search** is looking for something you expect to exist
  - Information tagging
  - Indexed search technologies like Apache Solr or ElasticSearch
  - The solution is pretty straightforward
- **Discovery** is finding something new, or in a new way
  - This is non-trivial
  - Traditional ontological method doesn't quite add up
  - The strength of semantic web is in inference
  - Need method involves
    - Dynamic data ranking
    - Dynamic update to the ontology
    - Mining user interaction and news outlets
- **Relevancy** is
  - Domain-specific
  - Personal
  - Temporal
  - Dynamic



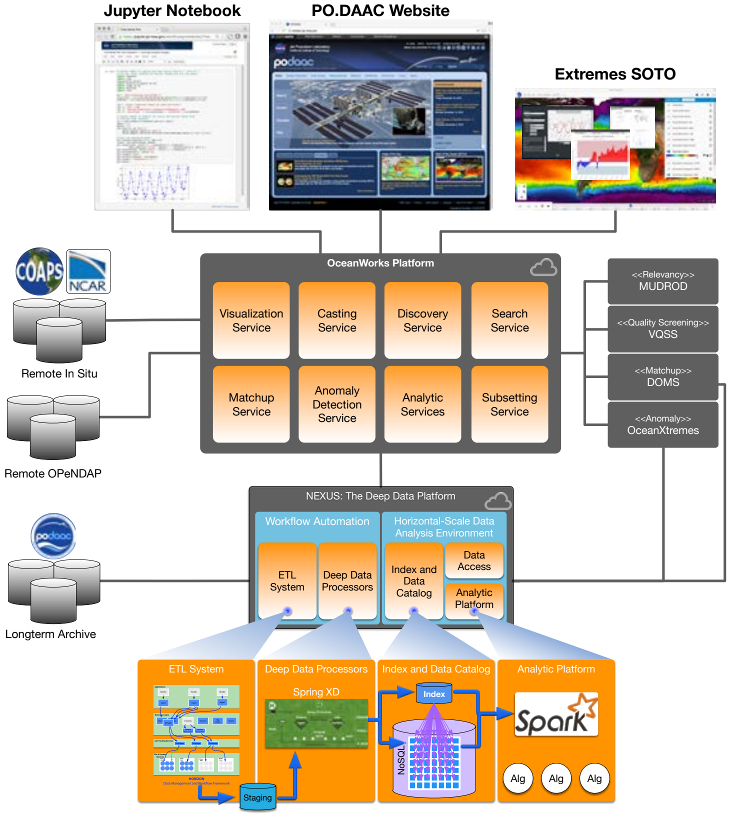
National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

AIST OceanWorks

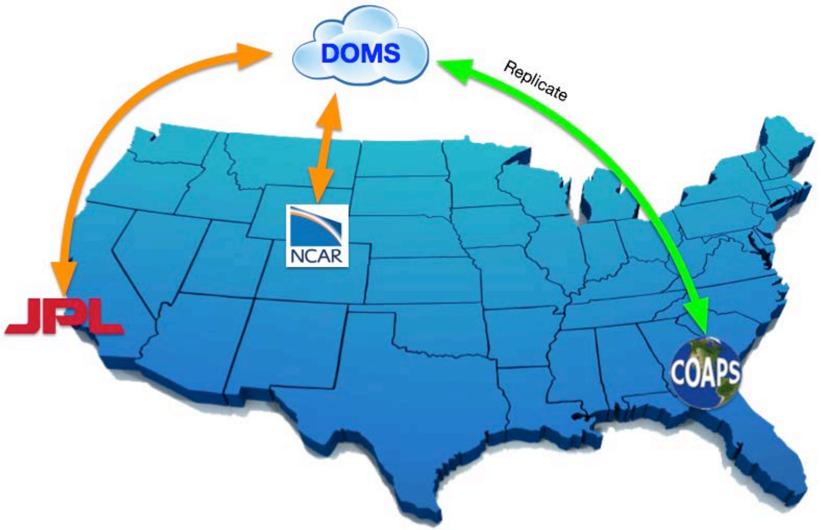
PI: Thomas Huang

- **OceanWorks** is to establish an **Integrated Data Analytic Center** at the NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC) for Big Ocean Science
- Focuses on technology integration, advancement and maturity
- Collaboration between JPL, FSU, NCAR, and GMU
- Bringing together PO.DAAC-related big data technologies
    - Anomaly detection and ocean science
    - Big data analytic platform
    - Distributed in-situ to satellite matchup
    - Search relevancy and discovery – linking datasets, services, and anomalies through recommendations
    - Metadata translation and services aggregation
    - Fast data subsetting
    - Virtualized Quality Screening Service



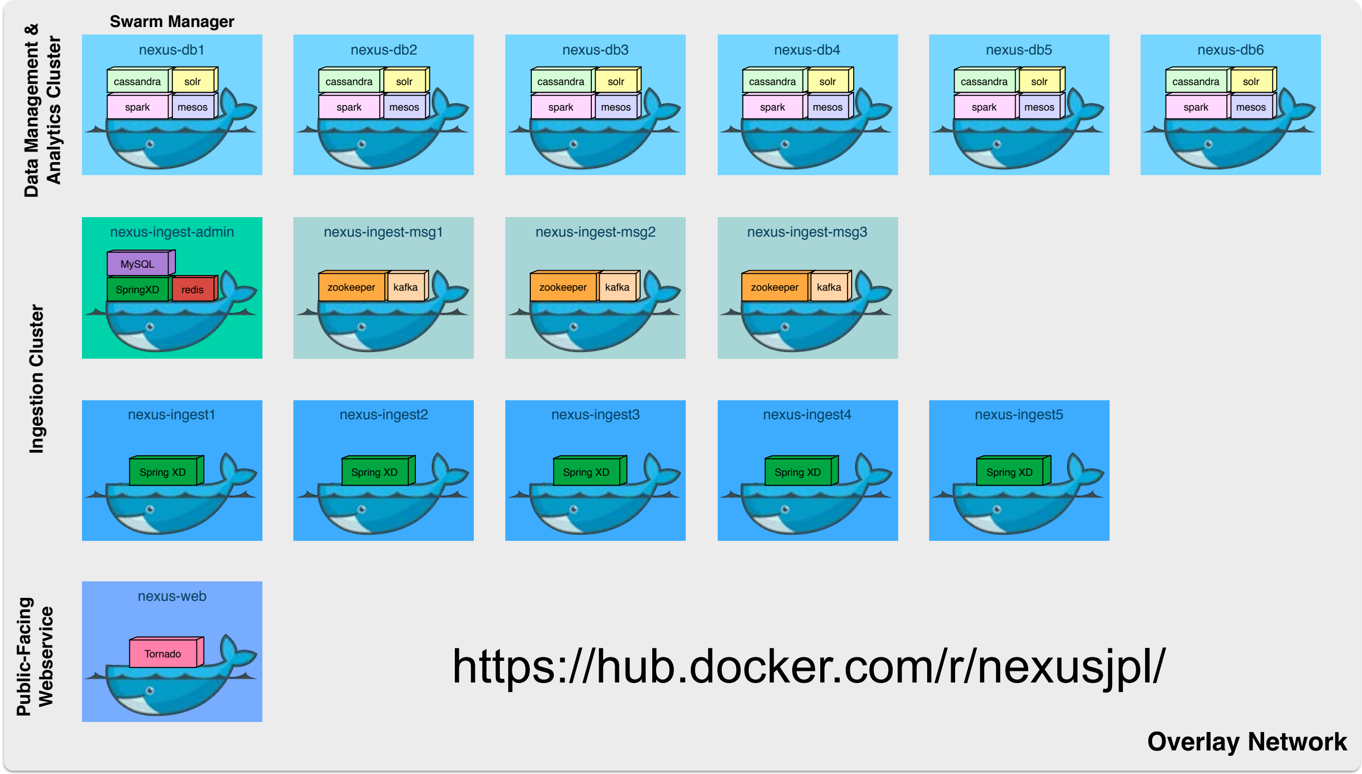
In Situ Data Replication

- DOMS now provides solution for replicating remote in situ data. Why?
    - AIST-14 DOMS demonstrated the ability to perform on-the-fly matchup by subsetting remote in situ measurements
    - Subsetting SAMOS data in FSU is a major performance bottleneck, due to
    - CONUS data transfer
    - Unstable network connections
- Replication is a background service to continue harvesting new SAMOS data into an Amazon Solr service
- Initial benchmarking shows significant matchup speed improvement
- Given access to in situ data is URL-based and using Open Search, there is not code changes to DOMS



Container-based deployment using Docker

- Dockerized all service components
- Register with DockerHub
- Integration with Continuous Integration (CI) solutions (e.g. Jenkins or Atlassian Bamboo)
- Same container deployment process for on-premise hardware or commercial Cloud environment

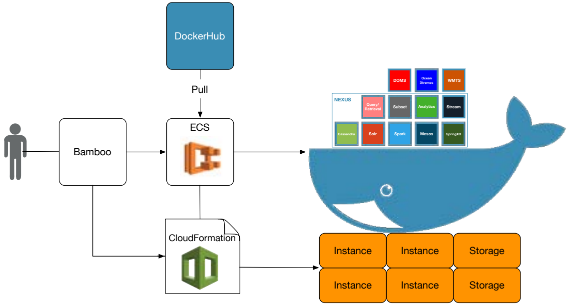


Example Container Deployment of our Open Source Analytic Engine, NEXUS

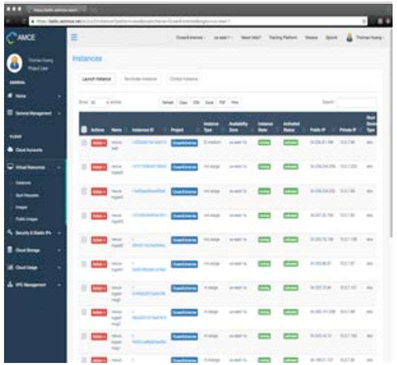
Deployment Automation

- Cloud Deployment is nontrivial
- Infrastructure Definition
    - Various machine instances
    - Storage and buckets
- Software Deployment.. manually
    - Build
    - Package
    - Install
    - Configure
    - Shell login (security issues)
- Best Practice: Deployment Automation
    - Script Infrastructure Definition (e.g. Amazon CloudFormation)
    - Container-based Deployment (e.g. Amazon ECS and DockerHub)
    - Eliminate SSH access to instances

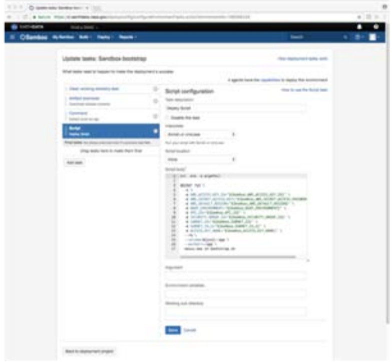
Infrastructure Definition Automation



Container Deployment Automation



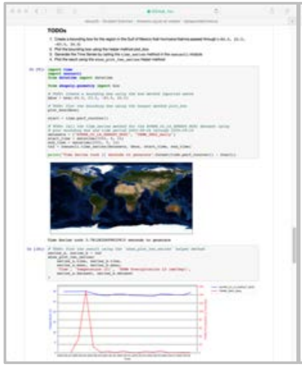
AMCE Deployment

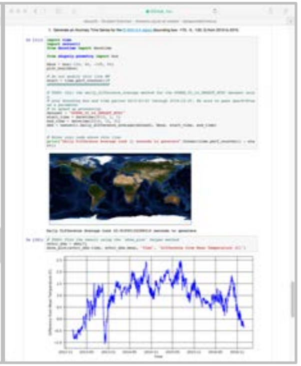


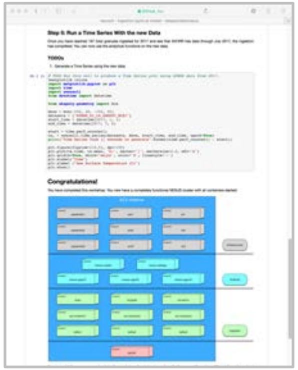
NGAP Deployment

Community Engagement and Support

- Develop in the open
- Working with Apache Incubator
- Target Apache top-level project by 2019.
- Public hands-on workshops
- Organize technical sessions at conferences
- Invited speaker and panelist
- Lead Editor: 2018 Wiley Book on **Big Earth Data Analytics in Earth, Atmospheric and Ocean Sciences**







Analyze Hurricane Katrina by comparing SST and TRMM time series

Generate daily difference average
“The Blob” is an oceanographic anomaly

Each participant deployed 3 computing clusters, a total of 24 containers on EC2

In Summary

- Traditional method for scientific research (search, download, local number crunching) is unable to keep up
- How much speed and storage can you afford?
- Think beyond archive and file downloads
- Investment in data and computational sciences
- Data Centers might want to be in the business of Enabling Science!
- Connected information enables discovery
- Community developed solution through open sourcing
- Thanks to the NASA ESTO/AIST and Sea Level Rise programs, and the NASA ESDIS project
- OceanWorks infusion 2018 – 2019
  - Watch for changes to the Sea Level Change Portal
    - Even faster analysis capabilities
    - More variety of measurements – satellites, in situ, and models
    - Event more relevant recommendations
  - NASA's Physical Oceanography Distributed Active Archive Center (PO.DAAC)
    - More than just pretty pictures. SOTO will have new analytic capabilities.
- Lead Editor: 2018 Wiley Book on **Big Earth Data Analytics in Earth, Atmospheric and Ocean Sciences**

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Thomas Huang
thomas.huang@jpl.caltech.edu

Jet Propulsion Laboratory
California Institute of Technology

BACKUP

Supported Datasets

- **Atmosphere**
  - MODIS Aqua Daily L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) (MOD08_D3v6)
  - MODIS Terra Daily L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) MOD08_D3v6)
  - MODIS Aqua Monthly L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) (MOD08_D3v6)
  - MODIS Terra Monthly L3 Atmospheres, Collection 6, variable Aerosol Optical Depth 550 nm (Dark Target) MOD08_D3v6)
- **Chlorophyll**
  - MODIS Aqua Level 3 Global Daily Mapped 4 km Chlorophyll a
- **Estimating the Circulation and Climate of the Ocean (ECCO)**
  - Monthly Mean Version 4 release 2 – Net Surface Fresh-Water Flux, Net Surface Heat Flux, Mixed-Layer Depth, Bottom Pressure, SEAICE Fractional Ice-Covered Area, Free Surface Height Anomaly, SEAICE Effective Snow Thickness, Total Heat Flux, Total Salt Flux
  - Monthly Mean Version 4 release 1 – Net Surface Fresh-Water Flux, Net Surface Heat Flux, Mixed-Layer Depth, Ocean Bottom Pressure, SEAICE Fractional Ice-Covered Area, Free Surface Height Anomaly, SEAICE Effective Snow Thickness, Actual Sublimation Freshwater Flux, Total Heat Flux, Total Salt Flux
- **Gravity**
  - Center for Space Research (CSR) GRACE RL05 Mascon Solutions
  - JPL GRACE Mascon Ocean, Ice, and Hydrology Equivalent Water Height RL05M.1 CRI filtered Version 2
- **Ocean Temperature**
  - GHRSST Level 4 MUR Global Foundation Sea Surface Temperature Analysis (v4.1)
  - GHRSST Level 4 AVHRR_OI Global Blended Sea Surface Temperature Analysis (GDS version 2) from NCEI
  - MODIS Aqua Level 3 SST Thermal IR Daily 4km Nighttime v2014.0
  - MODIS Aqua Level 3 SST Thermal IR Daily 4km Daytime v2014.0

Supported Datasets (+)

- **Salinity**
  - JPL SMAP Level 2B CAP Sea Surface Salinity V2.0 Validated Dataset
  - JPL SMAP Level 3 CAP Sea Surface Salinity Standard Mapped Image Monthly V3.0 Validated Dataset
- **Sea Surface Height Anomalies (SSHA)**
  - JPL MEaSUREs Gridded Sea Surface Height Anomalies Version 1609
- **Wind**
  - Cross-Calibrated Multi-Platform Ocean Surface Wind Vector L3.0 First-Look Analyses
- **Precipitation (non-ocean data)**
  - TRMM (TMPA) Precipitation L3 1 day 0.25 degree x 0.25 degree V7 (TRMM_3B42_Daily) at GES DIS
  - TRMM (TMPA-RT) Precipitation L3 1 day 0.25 degree x 0.25 degree V7 (TRMM_3B42_RT) at GES DISC
- **In Situ**
  - Shipboard Automated Meteorological and Oceanographic System (SAMOS)
  - International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 3, Individual Observations
  - Salinity Process in the Upper Ocean Regional Study – 1 (SPURS1)
  - Salinity Process in the Upper Ocean Regional Study – 2 (SPURS2)
  - Global gridded NetCDF Argo only dataset produced by optimal interpolation (salinity variables)
  - Global gridded NetCDF Argo only dataset produced by optimal interpolation (temperature variables)